I/S: A JOURNAL OF LAW AND POLICY FOR THE INFORMATION SOCIETY

Big Data Open Standards and Benchmarking To Foster Innovation

ASHIT TALUKDER, PHD*

Abstract: Big Data is a cross-cutting area that has the potential to improve decision making and yield better insight from data. While it is evident that Big Data is already resulting in improved data usage and has an increased role in decision-making across multiple domains, the rapid growth of Big Data technologies by multiple stakeholders is resulting in solutions that often cannot be adequately measured and characterized, or interoperate with other solutions. The potential of Big Data access, usability, information discovery and analytics, and management of data across federal and commercial sectors can be improved through open standards, measurements, evaluations, benchmarking, reference datasets, and development of reference frameworks.

I. INTRODUCTION

Data collected and analyzed from multiple sources like social media, financial transactions, communication systems, scientific experiments and sensors is increasing rapidly in volume and complexity. This new paradigm is creating a transformational shift in the way decisions are made and knowledge is extracted. This has been driven partly by rapid advances in wired and wireless networking and storage capabilities, computing infrastructure (scalable hardware and middleware designs), wider availability of open data, better search and information retrieval tools, and also cultural shifts toward a more persistent online presence of organizations and individuals.

* Program Director For Data Science, National Institute of Standards and Technology.

The persistent presence of data and wide-ranging impact of the inherent information contained in the data have led to the emergence of a new field called Big Data and Data Science. Big Data is an area that is impacting several segments of the U.S. economy including the sciences, engineering, manufacturing, government, and healthcare, to name a few. Both the Office of Management and Budget and the White House Office of Science and Technology Policy have highlighted Big Data as one of the federal government's multiagency science and technology priorities. The federal government has announced new investments of more than $200M in Big Data research across several federal agencies:

- The Defense Advanced Research Project Agency (DARPA) has funded technologies for the next generation of Big Data management, analytics, and visualization in the DARPA XDATA program through the I2O program office.[1]

- The National Institute of Health (NIH) has recognized the important impact of data science to the future of the healthcare and medical communities and has formed a Big Data to Knowledge (BD2K) initiative cutting across all the NIH institutes.[2] One of the goals of the NIH BD2K Initiative is to make NIH-funded research results and data accessible to anyone using new indexing, tagging and search and query tools, based on standards and interoperability.

- The National Science Foundation (NSF) has several programs focused on Big Data challenges such as data management, infrastructure and engineering at the Directorate for Computer and Information Science and Engineering (CISE), Data Infrastructure Building Blocks (DIBBS), and Computational and Data Enabled Science and Engineering.

- The Department of Energy (DOE) Office of Science formulated an Exascale Roadmap to highlight the areas of focus at DOE to advance Big Data capabilities.

---

[1] "XDATA," *Defense Advanced Research Projects Agency*, http://www.darpa.mil/Our_Work/I2O/Programs/XDATA.aspx.

[2] "NIH Big Data to Knowledge (BD2K)," *National Insitutes of Health*, http://bd2k.nih.gov/#sthash.Wmva4ZON.dpbs.

The importance of Big Data in operational environments is further evidenced in the number of deployments of Big Data solutions across sectors including finance, commerce, e-commerce, healthcare and security, and also the number of available Big Data commercial tools. Recognizing the value of publicly accessible data and the power of citizen science through data, a number of governments have started open data initiatives where data available at different federal, state and local agencies is made openly and freely available. A number of datasets (currently over 130,000 datasets) from the U.S. are available at data.gov ranging from agriculture and climate, to energy, education, and finance.

The impact is hardly limited to the United States. Per the well-cited report on Big Data by McKinsey,[3] improving and enhancing Big Data usage in the European public sector administration alone will add €250 billion value per year translating to 0.5 percent annual productivity growth. Data analytics within data science offers increasing capacity for enterprises and governments to analyze and use information. Analytics in data science has tremendous job creation potential. By 2015, according to Gartner,[4] data-driven innovation is projected to help create 4.4 million IT jobs globally, of which 1.9 million will be in the U.S.

II. GAPS AND CHALLENGES IN BIG DATA

Big Data has the potential to transform various sectors ranging from scientific discovery to medicine and healthcare to defense. Open data is being released by organizations to enable individuals and organizations to utilize such data for their applications. However, studies have shown that available data is often not easily accessible and usable, which lowers the utilization of data and underlying knowledge. While several governments and agencies have created Open Data Initiatives, studies have shown[5] that amongst all such data

---

[3] James Mankiya, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh and Angea Hung Byers, "Big Data: The Next Frontier for Innovation, Competition, and Productivity, " *McKinsey Global Institute*, May, 2011, http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.

[4] Patrick Thibodeau, "Gartner: Big Data to Create 1.9M IT Jobs in U.S. by 2015," *InfoWorld*, Oct. 22, 2012, http://www.infoworld.com/d/big-data/gartner-big-data-create-19m-it-jobs-in-us-2015-205417.

[5] "The Open Data Economy: Unlocking Economic Value by Opening Government and Public Data," *Capgemini Consulting*, http://www.capgemini.com/resources/the-open-data-economy-unlocking-economic-value-by-opening-government-and-public-data.

initiatives, over 60% lack enhanced search capabilities, and over 87% have negligible or minimal user participation.

These deficiencies point to the paucity of, and need for, standards and interoperability. Although the availability of data has resulted in an increase in the number of analytics and search solutions, little is known about which classes of analytics work for which sets of problems and data-types, and how the performance of such solutions compare to other existing or new tools. Rigorous measurements and benchmarking of data search, retrieval and analytics tools could be used to help improve the accuracy, performance and usability of data science solutions and characterize which classes of analytics are more suitable for specific data types and problem sets. In addition, there is currently only limited coordination between experts and scientists across sectors and domains which limits knowledge exchange and technology sharing. There is also a very limited number of analytics solutions that work robustly with multimodal and heterogeneous data types.

III. OPEN STANDARDS AND REFERENCE FRAMEWORKS FOR BIG DATA

Standards are an important element to ensure that Big Data systems and subcomponents can interoperate with each other. Open standards are a useful "glue" that can be utilized by multiple systems and technologies, legacy or new, and by components in Big Data systems to exchange information and data and work cohesively together. Several initiatives and working groups have been formed to address standards for Big Data. ISO/IEC JTC1's data management and interchange standards committee (SC32) has initiated a study on next generation analytics and Big Data. W3C has created several community groups on different aspects of Big Data such as analytics, data management, and others. In addition, a variety of existing;[6] software standards support the interoperability of data analytics for unstructured information, ontologies for information models, predictive models, and rules for specific applications. These include:

- RDF (Resource Description Framework), a standard data interchange model on the Web that operates even if the underlying schemas differ);[7]

[6] Certain commercial software, tools, and solutions are identified in this paper to foster understanding. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the software or equipment identified are necessarily the best available for the purpose.

[7] "Resource Description Framework," *W3C Semantic Web*, http://www.w3.org/RDF/.

- OWL (Web Ontology Language), a semantic web language designed to represent rich and complex knowledge about things, groups of things, and relations between things;[8]

- UIMA (Unstructured Information Management Applications), an Apache framework and tools to facilitate analysis of, and to create structure and knowledge from, unstructured content such as text, audio and video);

- PMML (Predictive Model Markup Language), a standard to allow interoperability between different closed-source implementations of machine learning and data mining algorithms and tools; and,

- RIFF (Resource Interchange File Format), a generic file container format for storing data in tagged chunks.

Machine readable data is another branch of standards-oriented data-activities. Persistent Identifiers (PIDs)[9] and Digital Object Identifiers (DOI)[10] are examples of solutions that facilitate and improve access to machine-readable data. DOIs are compilations of identifier registries that make a collection of identifiers actionable and interoperable. The DOI collection in turn can include identifiers from many other controlled collections or registries. DOIs can have varying scope and different levels of granularity for a variety of digital content types.

In addition to standards for Big Data, common taxonomies, reference frameworks and guidelines that address different aspects of Big Data systems would benefit the community immensely. A reference architecture is a resource containing a consistent set of architectural best practices for use by all the stakeholders of developers, designers and end-users. The idea behind reference architectures is to make a level playing field for an emerging or existing technology that may have several vendor implementations. A reference architecture enables standards to assist the consumers, end-users, and procurement officials to choose their vendors. For example,

---

[8] "Web Ontology Language, "*W3C Semantic Web*, http://www.w3.org/2001/sw/wiki/OWL.

[9] "USGS Data Management," *USGS Data Management*, http://www.usgs.gov/datamanagement/preserve/persistentIDs.php.

[10] "DOI Handbook," *DOI*, http://www.doi.org/hb.html.

the reference architecture for the cloud serves as a standard that provides confidence to consumers that their applications will work in any cloud. The NIST Cloud reference architecture[11] is more focused on the functional components ("what is needed") rather than a specific cloud implementation. The utility of reference architectures is evident from the Cloud Reference Architecture that was developed by NIST in collaboration with public and private stakeholders. The NIST Cloud security reference and risk management framework has resulted in the Federal Risk and Authorization Management Program[12] (FedRAMP), a government-wide program that provides a standardized approach to security assessment, authorization, and continuous monitoring for cloud products and services. FedRAMP is consistently used in many cloud specifications and requirements.

A number of parallel efforts have been undertaken in Big Data Reference frameworks. NIST has formed a Public Working Group (PWG) with associated subgroups to define Big Data taxonomies, reference architectures, security and privacy, and a Big Data technology roadmap. The details of the NIST Big Data PWG charter and reports from this PWG are available at http://bigdatawg.nist.gov/home.php. The NIST Big Data PWG has formulated a definition for Big Data, common taxonomies that can be used across multiple stakeholder communities, and a Big Data reference architecture that is a representation of a vendor-neutral and technology agnostic system, and a functional architecture that comprises logical roles and is applicable to a variety of business models. A few vendors,[13] including IBM, Oracle, SAP, and Pivotal have designed Big Data Reference architectures to which their Big Data solutions and implementations are mapped. In addition, the Cloud Security Alliance (https://cloudsecurityalliance.org/) has established a Big Data working group to identify solutions for data-centric security and privacy problems.

---

[11] Fang Liu, Jin Tong, Jian Mao, Robert Bohn, John Messina, Lee Badger and Dawn Leaf, "NIST Cloud Computing Reference Architecture, " *NIST Special Publication* 500-292, (2012).

[12] Matt Goodrich, "Federal Risk and Authorization Management Program," *GSA*, June, 2013, http://csrc.nist.gov/ groups/SMA/forum/ documents/june2013_presentations/ forum_june2013_mgoodrich.pdf.

[13] Certain commercial vendors and solutions are identified in this paper to foster understanding. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the vendors or their solutions identified are necessarily the best available for the purpose.

IV. OPEN MEASUREMENTS AND EVALUATIONS FOR BIG DATA SCIENCE

Big Data science and analytics promise to improve the quality of decisions and knowledge by moving from the previous paradigm of hypothesis-driven discovery to a new paradigm of data-driven discovery. Data-driven innovation could yield new knowledge hidden in the multitude of data based on hypotheses that did not previously exist. This could occur through finding previously unknown correlations, causation effects, detection of data outliers that could not be detected earlier due to the lack of sufficient sample sizes, or other pattern discovery techniques that could yield new potential hypotheses which might be further validated by more experimentation and analysis. Since data science is driving crucial decision-making, it is critical to understand the approaches, measure the performance of the underlying technologies, and to correctly interpret the final output.

There are, however, a number of challenges in the understanding and measurements of Big Data analytics solutions, of which the most salient are:

- Lack of understanding of what works (and does not work) in Big Data analytics algorithms;

- Lack of objective understanding of the foundational gaps in Big Data science (for example, what is needed to achieve robust analytics on distributed datasets that reside in geographically disparate locations, or foundational methodologies that can analyze and aggregate information from multiple data types such as text, images, video and transaction log-data?);
- Lack of accepted evaluation methods, tools, and reference data in Big Data;

- Lack of understanding regarding the usability of Big Data systems and solutions;

- Limited understanding of uncertainty propagation in Big Data systems and how the quality and context of the input data affect the resulting discoveries and derived conclusions (for example, how do noise and statistical uncertainty in input data propagate through a Big Data system, and how can uncertainties in the decision making process by a human-in-the-loop be quantified?);

- A lack of multidimensional benchmarks that can be applied to the analytics tools and processes; and,

- A way of evaluating which components are best suited for specific families of tasks.

Big Data solutions often involve a variety of data types and scalable distributed processing solutions, utilizing complex workflows comprising connected components and tools and human-in-the-loop. The complexity of a Big Data system could be many-fold at all functional components: during data capture and storage, data transformation, analytics, or at the human-system interface. Data being ingested by a Big Data system could be archived or at rest, or in some cases, the data could be streaming in real-time. Data could take various forms, as structured data residing in traditional databases in row-column format, as temporal signals, or in semi-structured or unstructured formats such as natural text, images, and videos. The variety of input data calls for increased complexity in the transformation and analytics processes utilized to yield insightful knowledge from the data. The analytics tools could be of various forms depending on data volume and variety, and the noise inherent in the captured data. Analytics typically involve techniques to clean up the data, handle missing data samples, and reduce noise in the data. In many cases, preparing the data for analytics takes more time than the analytics process itself.[14] The human-system interaction involving visualization or advanced interfaces of the analytics results is a critical piece that drives the decision making by the human-in-the-loop. Visualization tools need to present the right information at the right time in the right format for the human to effectively make the right decision in a timely manner out of the large volumes and streams of information. It is obvious that a workflow of a typical Big Data system is complex and non-trivial, and therefore measuring a Big Data system is also quite complex.

A more comprehensive understanding of the efficacy of Big Data solutions can be obtained if Big Data solutions can be measured through multiple parametric specifications such as accuracy, speed, resource utilization, network throughput, scalability, generalizability, and usability, among others. In addition, the Big Data community would benefit tremendously from the use of common reference datasets and open community challenge problems, such as the text

[14] Thomas Davenport, "Analytics 3.0," *Harvard Business Review*, December 2013.

search and retrieval evaluations and challenge problems used at the Text Retrieval Conference (TREC),[15] with specific tasks that would allow the benchmarking of different Big Data solutions along various parametric dimensions, and allow the identification of foundational gaps in Big Data science. Identification of foundational gaps would enable communities of stakeholders to address, work collaboratively, and thereby overcome such gaps. The text search and retrieval community has benefited immensely from the TREC design and development of open reference text datasets, including the formulation of metrics and open challenge problems with large research community involvement for specific tasks to help advance text search and information retrieval capabilities. TREC was formed by NIST in 1992, as a follow-up to a DARPA program, to solve two major problems in information retrieval. There were no data sets, i.e., document collections, with which to test information retrieval systems and techniques. At that time, there was burgeoning research in information retrieval algorithms and systems, but no metrics or methodologies to facilitate the standardized comparison of IR systems. A lack of standard evaluation methodologies resulted in duplicative research and a lack of understanding about what the foundational gaps were in information retrieval. TREC offered to overcome these two major problems and thereby advance the state of art in information retrieval. According to Google's chief economist, Hal Varian, TREC "revitalized research on information retrieval."[16] In his judgment, "the yearly TREC conference fostered collaboration, innovation, and a measured dose of competition (and bragging rights) that led to better information retrieval." An economic impact assessment study of TREC was done by RTI international[17] that highlights the benefits and advances in research that were enabled by NIST's TREC evaluations and reference datasets. Similar impact has resulted from the open video search and retrieval annual challenge conferences, TRECVID, where reference video data collections and targeted challenge problems with associated scientific metrics resulted in progressive advances in video understanding, search and retrieval capabilities over the past few years.

---

[15] Ellen M. Voorhees and Donna Harman, "Overview of the Ninth Text REtrieval Conference (TREC-9)," *TREC*, 2000, -trec.nist.gov.

[16] Hal Varian, "Why Data Matters," *Google*, http://googleblog.blogspot.com/2008/03/why-data-matters.html.

[17] Economic Impact Assessment of NIST's Text Retrieval Conference (TREC) Program, *RTI International*, (2010) http://www.nist.gov/director/planning/upload/report10-1.pdf.

Based on the previous successes and impact of the TREC conference to advance capabilities in text search and retrieval, and TRECVID to advance video understanding capabilities, a Big Data science and analytics challenge problem with reference datasets would enable similar foundational and applied improvements in Big Data analytics research and development capabilities. The shaping and grounding of future Big Data science R&D directions could be achieved through:

- The advancement of rigorous measurement techniques;
- The development of reference frameworks and reference datasets;
- The development of open challenge problems on use cases addressing compelling classes of technology challenges; and,
- Community collaboration (engaging stakeholders from all sectors).

Such open community engagement through reference data and challenge tasks would benefit a wide range of stakeholders in the Big Data community. These would include:

- End Users, who will have an objective resource to understand the technology space and tools best suited for their domains;
- Industry, which will have a means for engaging a broad R&D community to understand the state-of-the-art and best practices for implementing, developing and integrating appropriate tools and technology; and,
- Academia, Researchers and System Developers, who will have resources and methods for objectively identifying gaps, posing appropriate novel solutions by performing targeted research, and improving performance from both a component- and system-level perspective

V. CONCLUSION

In summary, while Big Data offers the scope for improved innovation in many sectors, numerous challenges remain before its potential can be realized. Technical challenges remain at many levels to improve the efficacy of these solutions. Rigorous measurements, testing, evaluations involving stakeholders from multiple sectors, and the adoption of open standards and reference frameworks could result in foundational improvements in Big Data science and analytics, data management, indexing search and query capabilities, knowledge discovery, information understanding and visualization.